Requested Patent: JP2001282628A

Title:

METHOD FOR DUPLICATING DATA OF STORAGE SUBSYSTEM AND DATA DUPLICATING SYSTEM ;

Abstracted Patent: EP1150210 ;

Publication Date: 2001-10-31 ;

Inventor(s): NAKANO TOSHIO (JP); TABUCHI HIDEO (JP); SHIMADA AKINOBU (JP) ;

Applicant(s): HITACHI LTD (JP) ;

Application Number: EP20000119133 20000904 ;

Priority Number(s): JP20000101168 20000331 ;

IPC Classification: G06F11/14; G06F11/20 ;

Equivalents: ;

ABSTRACT:

In order to realize asynchronous type system assuring the consistency of data with the function of disk subsystems without the need of introducing new software to a host unit and without the deterioration of the performance of a main center, in a remote copy system which copies the data to the disk subsystems of the remote center for duplicating the data in the disk subsystems of the main center, the disk subsystems give serial numbers and times to the data together with writing said data to the storage devices in the disk subsystem and transfer said data to the other disk subsystems, and the other disk subsystems arrange the two or more data in the sequence of the serial numbers, decide the oldest time among the latest time given to each of the disk subsystems communicating among the disk subsystems and the data given with the time not later than the decided oldest time are the objects of writing to each of the disk storage devices.

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号
特開2001－282628
（P2001－282628A）
(43)公開日　平成13年10月12日(2001.10.12)

| (51)Int.Cl.[7] | | 識別記号 | FI | | テーマコード(参考) |
|---|---|---|---|---|---|
| G06F | 12/16 | 310 | G06F | 12/16 | 310J　5B018 |
| | 3/06 | 304 | | 3/06 | 304E　5B034 |
| | 11/20 | 310 | | 11/20 | 310C　5B065 |
| | 12/00 | 533 | | 12/00 | 533J　5B082 |
| | 13/00 | 301 | | 13/00 | 301P　5B083 |

審査請求　未請求　請求項の数21　ＯＬ　（全 14 頁）

(21)出願番号　特願2000－101168(P2000－101168)

(22)出願日　平成12年３月31日(2000.3.31)

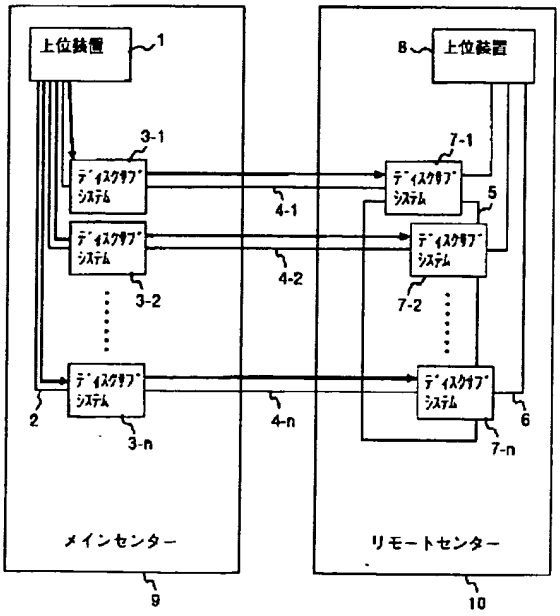
(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　田渕　英夫
神奈川県小田原市国府津2880番地　株式会社日立製作所ストレージシステム事業部内
(72)発明者　島田　朗仲
神奈川県小田原市国府津2880番地　株式会社日立製作所ストレージシステム事業部内
(74)代理人　100068504
弁理士　小川　勝男　（外１名）

最終頁に続く

(54)【発明の名称】　記憶サブシステムのデータ二重化方法及びデータ二重化システム

(57)【要約】
【課題】　メインセンターのディスクサブシステムのデータを二重化するためにリモートセンターのディスクサブシステムにコピーするリモートコピーシステムにおいて、上位装置に新規ソフトウェアの導入を必要とせず、ディスクサブシステムの機能のみでデータの整合性を保証し、メインセンターの性能低下のない非同期型のシステムを実現する。
【解決手段】　ディスクサブシステム３は、そのディスク記憶装置にデータ書き込みするとともに、該データに通番と時刻を付与し、ディスクサブシステム７へ転送する。ディスクサブシステム７は、受け取った複数のデータを通番の順に配列し、ディスクサブシステム７間の通信によって各ディスクサブシステム７についてそれぞれ付与された最新の時刻の中で最古の時刻を決定し、決定された最古の時刻以前の時刻が付与されたデータを各々のディスク記憶装置へのデータ書き込み対象とする。

図　1



【特許請求の範囲】

【請求項１】複数の記憶サブシステムから構成される第１の記憶サブシステム群と、第１の記憶サブシステム群のデータのコピーを記憶する複数の記憶サブシステムから構成される第２の記憶サブシステム群とを有するシステムのデータ二重化方法であって、第１の記憶サブシステム群に属する各記憶サブシステムによってその記憶装置にデータ書き込みするとともに、該データに通番と時刻を付与し、伝送路を介して第２の記憶サブシステム群に属する記憶サブシステムへ転送し、第２の記憶サブシステム群に属する各記憶サブシステムによって受け取った複数のデータを通番の順に配列し、第２の記憶サブシステム群に属する記憶サブシステム間の通信によって各記憶サブシステムについてそれぞれ付与された最新の時刻の中で最古の時刻を決定し、決定された最古の時刻以前の時刻が付与されたデータを各記憶サブシステムの記憶装置へのデータ書き込み対象とすることを特徴とする記憶サブシステムのデータ二重化方法。

【請求項２】前記第１の記憶サブシステム群に属する記憶サブシステムと前記第２の記憶サブシステム群に属する記憶サブシステムとを接続する前記伝送路は、ＳＡＮ（ストレージ・エリア・ネットワーク）を構成することを特徴とする請求項１記載の記憶サブシステムのデータ二重化方法。

【請求項３】前記第１の記憶サブシステム群に属する各記憶サブシステムによって前記時刻を参照するための時計を外部からの時刻情報によって補正することを特徴とする請求項１記載の記憶サブシステムのデータ二重化方法。

【請求項４】前記第２の記憶サブシステム群に属する記憶サブシステム間はループ伝送路によって接続されており、各記憶サブシステムによって自己の前記最新の時刻と受け取った前記最新の時刻のうちで古い方の時刻を隣の記憶サブシステムに伝送し、自記憶サブシステムが伝送しかつ自記憶サブシステムに戻った時刻を前記最古の時刻と決定することを特徴とする請求項１記載の記憶サブシステムのデータ二重化方法。

【請求項５】前記第２の記憶サブシステム群に属する複数の記憶サブシステムのうちの１つをマスタ記憶サブシステムに設定し、マスタ記憶サブシステム以外の各記憶サブシステムによって前記最新の時刻をマスタ記憶サブシステムに通知し、マスタ記憶サブシステムによって自己の最新の時刻と取得した最新の時刻の中から前記最古の時刻を決定することを特徴とする請求項１記載の記憶サブシステムのデータ二重化方法。

【請求項６】前記第１の記憶サブシステム群に属する複数の記憶サブシステムによって各々書き込みデータを前記第２の記憶サブシステム群に属する一の記憶サブシステムに転送し、第２の記憶サブシステム群に属する一の記憶サブシステムによって第１の記憶サブシステム群に属する前記記憶サブシステムごとに付与された最新の時刻を選定し、選定された最新の時刻の中で最古の時刻を前記最古の時刻の候補とすることを特徴とする請求項１記載の記憶サブシステムのデータ二重化方法。

【請求項７】前記記憶サブシステムの記憶装置は複数のボリュームによって構成され、コピー元である第１の記憶サブシステム群に属するボリュームとコピー先である第２の記憶サブシステム群に属するボリュームとがボリュームペアを構成するとき、第１の記憶サブシステム群に属する記憶サブシステムによって複数のボリュームペアから構成されるボリュームグループごとに第２の記憶サブシステム群へのデータ転送の開始及び停止を制御することを特徴とする請求項１記載の記憶サブシステムのデータ二重化方法。

【請求項８】複数の記憶サブシステムから構成される第１の記憶サブシステム群と、第１の記憶サブシステム群のデータのコピーを記憶する複数の記憶サブシステムから構成される第２の記憶サブシステム群とを有するデータ二重化システムであって、第１の記憶サブシステム群に属する各記憶サブシステムは、その記憶装置にデータを書き込みする手段と、該データに通番と時刻を付与する手段と、通番と時刻を付与された該データを伝送路を介して第２の記憶サブシステム群に属する記憶サブシステムへ転送する手段と、第２の記憶サブシステム群に属する各記憶サブシステムは、受け取った複数のデータを通番の順に配列する手段と、第２の記憶サブシステム群に属する記憶サブシステム間の通信によって各記憶サブシステムについてそれぞれ付与された最新の時刻の中で最古の時刻を決定する手段と、決定された最古の時刻以前の時刻が付与されたデータを各記憶サブシステムの記憶装置へデータ書き込みする手段とを有することを特徴とする記憶サブシステムのデータ二重化をするシステム。

【請求項９】前記第１の記憶サブシステム群に属する記憶サブシステムと前記第２の記憶サブシステム群に属する記憶サブシステムとを接続する前記伝送路は、ＳＡＮ（ストレージ・エリア・ネットワーク）を構成することを特徴とする請求項８記載の記憶サブシステムのデータ二重化をするシステム。

【請求項１０】前記第１の記憶サブシステム群に属する各記憶サブシステムは、前記時刻を参照するための時計を外部からの時刻情報によって補正する手段を有することを特徴とする請求項８記載の記憶サブシステムのデータ二重化をするシステム。

【請求項１１】前記第２の記憶サブシステム群に属する記憶サブシステム間はループ伝送路によって接続されており、各記憶サブシステムは、自己の前記最新の時刻と受け取った前記最新の時刻のうちで古い方の時刻を隣の記憶サブシステムに伝送し、自記憶サブシステムが伝送しかつ自記憶サブシステムに戻った時刻を前記最古の時

刻と決定する手段を有することを特徴とする請求項８記載の記憶サブシステムのデータ二重化をするシステム。

【請求項１２】前記第２の記憶サブシステム群に属する複数の記憶サブシステムのうちの１つがマスタ記憶サブシステムとして設定され、前記マスタ記憶サブシステム以外の各記憶サブシステムは、前記最新の時刻をマスタ記憶サブシステムに通知する手段を有し、前記マスタ記憶サブシステムは、自己の最新の時刻と取得した最新の時刻の中から前記最古の時刻を決定する手段を有することを特徴とする請求項８記載の記憶サブシステムのデータ二重化をするシステム。

【請求項１３】前記第１の記憶サブシステム群に属する複数の記憶サブシステムは、各々書き込みデータを前記第２の記憶サブシステム群に属する一の記憶サブシステムに転送するよう構成され、第２の記憶サブシステム群に属する前記一の記憶サブシステムは、第１の記憶サブシステム群に属する前記記憶サブシステムごとに付与された最新の時刻を選定する手段と、選定された最新の時刻の中で最古の時刻を前記最古の時刻の候補とする手段とを有することを特徴とする請求項８記載の記憶サブシステムのデータ二重化をするシステム。

【請求項１４】前記記憶サブシステムの記憶装置は複数のボリュームによって構成され、コピー元である第１の記憶サブシステム群に属するボリュームとコピー先である第２の記憶サブシステム群に属するボリュームとがボリュームペアを構成するとき、第１の記憶サブシステム群に属する記憶サブシステムは、複数のボリュームペアから構成されるボリュームグループごとに第２の記憶サブシステム群へのデータ転送の開始及び停止を制御する手段を有することを特徴とする請求項８記載の記憶サブシステムのデータ二重化をするシステム。

【請求項１５】複数の記憶サブシステムから構成される記憶サブシステム群に属する記憶サブシステムであって、前記記憶サブシステムは、外部から受け取ったデータをその記憶装置にデータ書き込みする第１の手段と、該データに通番と時刻を付与して他の記憶サブシステムに転送する第２の手段と、他の記憶サブシステムから受け取った複数のデータを前記通番の順に配列する第３の手段と、他の記憶サブシステム間の通信によって各記憶サブシステムについてそれぞれ付与された最新の時刻の中で最古の時刻を決定する第４の手段とを有し、当該記憶サブシステムがローカルモードのとき第１の手段と第２の手段を動作させ、リモートモードのとき第３の手段、第４の手段及び第１の手段を動作させ、決定された最古の時刻以前の時刻が付与されたデータを第１の手段によるデータ書き込み対象とすることを特徴とするデータ二重化システムを構成する記憶サブシステム。

【請求項１６】前記ローカルモードで動作する前記記憶サブシステムと前記リモートモードで動作する前記記憶サブシステムとは、ＳＡＮ（ストレージ・エリア・ネットワーク）を介して接続されることを特徴とする請求項１５記載のデータ二重化システムを構成する記憶サブシステム。

【請求項１７】前記記憶サブシステムは、前記時刻を参照するための時計を外部からの時刻情報によって補正する手段を有することを特徴とする請求項１５記載のデータ二重化システムを構成する記憶サブシステム。

【請求項１８】前記リモートモードで動作する前記記憶サブシステム間はループ伝送路によって接続されており、前記リモートモードの各記憶サブシステムは、自己の前記最新の時刻と受け取った前記最新の時刻のうちで古い方の時刻を隣の記憶サブシステムに伝送する手段と、自記憶サブシステムが伝送しかつ自記憶サブシステムに戻った時刻を前記最古の時刻と決定する手段とを有することを特徴とする請求項１５記載のデータ二重化システムを構成する記憶サブシステム。

【請求項１９】前記リモートモードで動作する前記記憶サブシステムのうちの１つがマスタ記憶サブシステムとして設定され、前記マスタ記憶サブシステム以外の各記憶サブシステムは、前記最新の時刻をマスタ記憶サブシステムに通知する手段を有し、前記マスタ記憶サブシステムは、自己の最新の時刻と取得した最新の時刻の中から前記最古の時刻を決定する手段とを有することを特徴とする請求項１５記載のデータ二重化システムを構成する記憶サブシステム。

【請求項２０】前記ローカルモードで動作する複数の記憶サブシステムは、前記リモートモードで動作する記憶サブシステム群に属する一の記憶サブシステムに該データを転送するよう構成され、前記リモートモードで動作する記憶サブシステム群に属する前記一の記憶サブシステムは、前記ローカルモードで動作する記憶サブシステムごとに付与された最新の時刻を選定する手段と、選定された最新の時刻の中で最古の時刻を前記最古の時刻の候補とする手段とを有することを特徴とする請求項１５記載のデータ二重化システムを構成する記憶サブシステム。

【請求項２１】前記記憶サブシステムの記憶装置は複数のボリュームによって構成され、コピー元である前記ローカルモードで動作する記憶サブシステムに属するボリュームとコピー先である前記リモートモードで動作する記憶サブシステムに属するボリュームとがボリュームペアを構成するとき、前記ローカルモードで動作する記憶サブシステムは、少なくとも１つのボリュームペアから構成されるボリュームグループごとに前記リモートモードで動作する記憶サブシステムへのデータ転送の開始及び停止を制御する手段を有することを特徴とする請求項１５記載のデータ二重化システムを構成する記憶サブシステム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、コンピュータが参照・更新するためのデータを格納する記憶サブシステムに係わり、特に記憶サブシステムの保有するデータを二重化する方法およびデータ二重化システムに関する。

【0002】

【従来の技術】地震等の災害に備えてコンピュータセンターとこれをバックアップするコンピュータセンターとを保有する会社、金融機関等が知られている。このようなバックアップ設備を有するシステムにおいて、メインのコンピュータセンターとリモートのコンピュータセンターとは、地理的に100km～数100km程度離れて設けられ、メインセンターとリモートセンターにそれぞれ設置されているディスクサブシステムの間ではデータは二重化して保有される。その方法の1つとしてメインセンター側のディスクサブシステムに発生した書き込みデータをリモートセンター側のディスクサブシステムへ転送し、同一データをリモート側のディスクサブシステムに書き込みする、いわゆるリモートコピー機能をもつシステムが既にいくつか実用化されている。リモートコピー機能は、同期型と非同期型の2種類に大別される。

【0003】同期型とはメインセンター内のホストコンピュータ（上位装置）からディスクサブシステムにデータの更新(書込み)指示が有ったとき、その指示対象がリモートコピー機能の対象でもあるときは、そのリモートコピー機能の対象であるリモートセンターのディスクサブシステムに対して、指示された更新（書込み）が終了してから、メインセンターの上位装置に更新処理の完了を報告する処理手順をいう。この場合、メインセンターとリモートセンターとの地理的距離に応じてこの間に介在するデータ伝送線路の能力の影響を受け、伝送時間等による時間遅れが発生する。

【0004】これに対し非同期型とは、メインセンター内の上位装置からディスクサブシステムにデータの更新(書込み)指示があったとき、その指示対象がリモートコピー機能の対象であっても、メインセンター内のディスクサブシステムの更新処理が終わり次第、上位装置に対し更新処理の完了を報告し、リモートセンターのディスクサブシステムに対するデータの更新（反映）をメインセンターに関する処理とは非同期に実行する処理手順をいう。非同期型によれば、メインセンター内部で必要とされる処理時間でデータ更新が終了するので、リモートセンターへのデータの格納に起因する時間遅れは発生しない。このため遠隔地間のディスクサブシステムに対してリモートコピーを行う場合、伝送時間等によるメインセンターの業務への影響を回避することを最優先に考えると、同期型のリモートコピーよりも非同期型のリモートコピーの方が適していると言える。

【0005】非同期型は、リモートセンターのディスクサブシステムのデータがメインセンター側のデータに対し常に一致しているわけではない。このためメインセンターが災害等により機能を失った場合は、リモートセンター側にデータの反映が完了していないデータが消失することとなる。しかしメインセンター側のディスクサブシステムのアクセス性能をリモートコピー機能を実施しない場合と同等レベルとすることができる。

【0006】かかる従来技術では上位装置が介在してリモートコピーの機能を達成するため、種々の課題があった。

【0007】（1）データの一貫性保全について

リモートコピーを行う場合、メインセンターのディスクサブシステムとリモートセンターのディスクサブシステム間は独立した通信リングで接続される。つまりメインセンターの複数のディスクサブシステムとリモートセンターの複数のディスクサブシステムの間でリモートコピーを行う場合は、独立した通信リンクで接続されたディスクサブシステムの組が複数存在する構成となる。複数のディスクサブシステムを有するメインセンターのバックアップをリモートセンターで行う場合、複数のディスクサブシステム間でデータの更新順序を保持する「一貫性保全」という問題がある。非同期型リモートコピーではリモートセンターへの更新データの反映がメインセンターでの実際の更新処理の発生時点より遅れて処理されることはやむを得ない。しかし更新の順序はメインセンターと一致していなければならない。あるいは更新の順序が常時一致していなくとも、少なくともユーザがリモートセンターでデータを必要とする時点において一貫性が保持されている状態のデータがリモートセンターに記録されていなければならない。

【0008】一般にデータベースは、データベース本体と、データ更新の履歴を記録する各種ログ情報および制御情報から構成されており、データベース本体と各種ログ情報および制御情報は、信頼性への配慮から異なるディスクサブシステムに記録されるようにシステム設計が行われることが多い。ただしデータベース本体と各種ログ情報および制御情報はそれぞれが関連性を持っており、データ更新の際はデータベース本体に加え、これらログ情報、制御情報をも追加・更新し、システムの整合性が保たれている。これら一連の更新が行われる時間間隔は、短い場合には数マイクロ秒程度の間隔で順次実行されることがある。仮に更新の順序が崩れた場合、更新順序に関連するこれらの情報の整合性も崩れ、最悪の場合には、データベース全体の破壊につながる可能性がある。

【0009】例えばメインセンターではデータベースの更新後にログ情報等の更新が行われたとしても、リモートコピーシステムを構成する上記の通信リンクの事情によっては更新されたログ情報等がデータベース本体の更新情報よりも先にリモートセンターに到着する可能性がある。このためにリモートセンターではログ情報等がデ

ータベース本体よりも先に追加・更新されるといった状況を生む可能性が潜んでいる。仮にログ情報等のみが追加・更新され、それに関連するデータベース本体がまだ更新されていない論理的に不整合な状態でメインセンターが被災した場合、リモートセンターのデータベース自体が有用なものとなり得なくなる。このためメインセンターでデータが更新された順序と同じ順序でリモートセンターにおいてもデータが更新されなければならないという課題があった。

【００１０】（２）上位装置が介在するリモートコピー機能について

メインセンター及びリモートセンターに複数のディスクサブシステムが存在する一般的な環境で非同期型のリモートコピーを実現する場合には、メインセンターの上位装置がディスクサブシステムへデータの更新を指示するとき、タイムスタンプなどの更新順序に関する情報をデータに付加し、このような時刻情報に基づいてリモートセンターの上位装置がそのディスクサブシステムの更新データ反映処理を実行する技術が知られている。例えば特開平６－２９０１２５号公報（米国特許第５４４６８７１号）が挙げられる。特開平６－２９０１２５号公報では、上位装置が介在してリモートコピー機能を実現している。具体的にはメインセンター側の上位装置のオペレーティングシステムとディスクサブシステム、リモートセンター側の上位装置のデータムーバーソフトウェアとディスクサブシステムの連携により、更新順序情報の発行、送付およびこれに基づく更新データの反映処理を実現している。

【００１１】

【発明が解決しようとする課題】上記従来技術によれば、メインセンターとリモートセンター間でデータの更新順序性を保証しながら非同期型のリモートコピー機能が実現できる。しかしこの技術は、上位装置とディスクサブシステムの双方に非同期型リモートコピー機能を実現するためのソフトウェアが必要であり、かつ両者が連携しなければならない。専用の新規ソフトウェアの導入が必要なため、ユーザはソフトウェアの導入、設定、検査、ＣＰＵ負荷増加に伴うシステム設計の見直し等の作業が発生する。このため従来技術の導入のためには所定の期間を要し、費用が発生するという導入障壁があった。またそもそも上位装置からタイムスタンプなどの更新順序に関する情報がデータに付加されないシステムや、複数のホストコンピュータの内部時計を合わせる機構を備えていないシステム、特にオープンシステムにおいてはその実現手段がないのが実情である。従って上位装置のタイプや上位ソフトウェアに係わらず、ディスクサブシステムの機能のみで非同期型のリモートコピー機能を実現するためには、ディスクサブシステムの機能のみでデータの更新順序の整合性を保持する必要がある。更新順序の整合性が必要なデータが複数のディスクサブシステムに分散されて格納されている場合、複数のディスクサブシステム間で更新順序の整合性を保持するための手段がないという課題もあった。

【００１２】本発明の目的は、上位装置に新規ソフトウェアの導入を必要とせず、記憶サブシステムの機能のみでデータの整合性を保証し、導入が容易かつメインセンターのコンピュータの性能低下の少ない非同期型のリモートコピー機能を実現することにある。

【００１３】

【課題を解決するための手段】本発明は、複数の記憶サブシステムから構成される第１の記憶サブシステム群と、第１の記憶サブシステム群のデータのコピーを記憶する複数の記憶サブシステムから構成される第２の記憶サブシステム群とを有するシステムのデータ二重化方法であって、第１の記憶サブシステム群に属する各記憶サブシステムによってその記憶装置にデータ書き込みするとともに、このデータに通番と時刻を付与し、伝送路を介して第２の記憶サブシステム群に属する記憶サブシステムへ転送し、第２の記憶サブシステム群に属する各記憶サブシステムによって受け取った複数のデータを通番の順に配列し、第２の記憶サブシステム群に属する記憶サブシステム間の通信によって各記憶サブシステムについてそれぞれ付与された最新の時刻の中で最古の時刻を決定し、決定された最古の時刻以前の時刻が付与されたデータを各記憶サブシステムの記憶装置へのデータ書き込み対象とする記憶サブシステムのデータ二重化方法を特徴とする。またこのような方法に従って機能するリモートコピーシステムを特徴とする。

【００１４】

【発明の実施の形態】以下、図面を参照しながら本実施形態のシステムについて説明する。

【００１５】図１は、各々コンピュータを装備する２つのデータセンターの間でデータの二重化を行うシステムの構成図である。メインセンター９側の複数台のディスクサブシステム３－１、３－２、……３－ｎと、リモートセンター１０側の複数台のディスクサブシステム７－１、７－２、……７－ｎは、上位装置１，８を介さずに互いに接続され、両センター間でデータの二重化を行うリモートコピーシステムを実現している。上位装置１，８を介さないディスクサブシステムの接続としては、ここでは詳細に記述しないが例えばＳＡＮ（ストレージ・エリア・ネットワーク：Storage Area Network）を利用した接続が挙げられる。

【００１６】メインセンター９において、データ処理を行う中央処理装置（ＣＰＵ）を持つ上位装置１は、インタフェースケーブル２を介してディスクサブシステム３－１、３－２、……３－ｎに接続されている。

【００１７】ディスクサブシステム３－１は、インタフェースケーブル４－１を介してリモートセンタのディスクサブシステム７－１と接続され、ディスクサブシステ

ム3－2はインタフェースケーブル4－2を介してリモートセンタのディスクサブシステム7－2と接続され、同様にディスクサブシステム3－nはインタフェースケーブル4－nを介してリモートセンタのディスクサブシステム7－nと接続される構成をとる。以下ディスクサブシステム3－1、3－2、……3－nのいずれか1台あるいはディスクサブシステム7－1、7－2、……7－nのいずれか1台を指すときには、ディスクサブシステム3あるいはディスクサブシステム7と呼ぶことがある。他の機構についても同様である。

【0018】なおセンター間を結ぶインタフェースケーブル4－1，4－2，……4－nには、LED駆動装置によって駆動されている光ファイバリンクや、光ファイバケーブルを用いて一般にファイバチャネルと呼ばれるインタフェースプロトコルで駆動されているものが含まれている。またT3ネットワークやATMネットワークに代表される電気通信リンクを含んでもよい。従ってメインセンター9のディスクサブシステム3と、リモートセンター10のディスクサブシステム7の間には、途中に一般のファイバチャネルスイッチを接続したり、T3ネットワークやSONETネットワーク及びATMネットワーク等と接続することも可能である。図1では特に図示しないが、本例ではこれらの接続形態が可能である点も含めてインタフェースケーブル4と称する。

【0019】リモートコピーの対象となるデータが格納されるリモートセンター10内の任意のディスクサブシステム7－1，7－2，7－nは、同センター内に存在しリモートコピーの対象となるデータが格納される他の一台のディスクサブシステムとインタフェースケーブル5を介して接続される。本実施形態では、ディスクサブシステム7－1はディスクサブシステム7－2と、ディスクサブシステム7－2はディスクサブシステム7－3と、ディスクサブシステム7－nはディスクサブシステム7－1といった様にディスクサブシステム7が順次接続されるループ構成をとる。

【0020】上位装置8は、リモートセンター10においてディスクサブシステム7－1、7－2、……7－nとインタフェースケーブル6によって接続され、ディスクサブシステム7－1、7－2、……7－nに対し、参照及び更新を行う中央処理装置である。上位装置8は、メインセンター9の上位装置1が災害や故障等により本来の機能を果たせなくなった場合に、上位装置1の代替となって処理を行うことができる。このほかディスクサブシステム7－1、7－2、……7－nに格納されているデータを使用して、メインセンター9の上位装置1とは異なる処理を上位装置1とは独立に実行することができる。ただし上位装置8がディスクサブシステム7－1、7－2、……7－nに対し処理を行わない場合には上位装置8は不要である。

【0021】図1に示すシステム構成は、上位装置1から複数台のディスクサブシステム3－1、3－2、……3－nに対しデータの書込み指示があった場合に、リモートセンター10内の複数台のディスクサブシステム7－1、7－2、……7－nにも同じデータがメインセンター9の処理とデータ更新の一貫性を保って格納されるシステム構成を示している。図1の矢印は、上位装置1から書込み指示のあったデータの流れを示している。

【0022】図2は、ディスクサブシステム3及びディスクサブシステム7の内部構成を示す図である。ディスクサブシステム3は、上位装置1から転送されたコマンド、データなどを授受したり、他のディスクサブシステム7と接続するためのインタフェース制御部11、上位装置1から参照又は更新されるデータを格納するキャッシュメモリ12、このデータを記録する記録媒体としての磁気ディスクドライブ13、このデータの管理情報、リモートコピーの状態情報、時刻情報などを格納する制御メモリ14、およびこれらの各要素を制御するディスクサブシステム制御部17を備える。ディスクサブシステム制御部17は、マイクロプロセッサ使用のプログラム制御によって動作する。

【0023】ディスクサブシステム7も同様の内部構成を有する。なおディスクサブシステム3とディスクサブシステム7とが同一の内部構成と機能を有するディスクサブシステムとし、そのディスクサブシステムがローカルモードのときディスクサブシステム3として動作し、リモートモードのときディスクサブシステム7として動作するようにしてもよい。ローカルモード及びリモートモードのとき各々動作する機能については、以下に説明される通りである。また一般に磁気ディスクドライブ13に相当する部分が磁気ディスク以外の記憶装置であり、ディスクサブシステム3及び7が広く記憶サブシステムと呼ばれるものであってもよい。

【0024】また、ディスクサブシステム3は、各々時計15と時計補正部16を備えている。時計補正部16は、各ディスクサブシステム3の装置筐体内もしくは各ディスクサブシステムとは別の位置で十分近くに存在するタイムサーバが配信する時刻情報をもとに時計15の補正を随時行い、タイムサーバが配信する時刻と各ディスクサブシステム3が保有する時計15の時刻との誤差を数マイクロ秒程度の範囲内におさめている。ここでタイムサーバとは、GPSや電波時計等の様に時刻情報が含まれた電波を受信する機能と、受信した時刻情報を各ディスクサブシステム3に伝達配信する機能を持つものを指す。なお電波受信機能がなくともタイムサーバ自身で時刻を生成する機能と時刻を各ディスクサブシステムに伝達配信する機能を有しているものであっても良い。またディスクサブシステム3のうちの1台がタイムサーバの機能を保有したものであれば、独立したタイムサーバは不要である。図2に示すタイムサーバ20は、ディ

スクサブシステム３の十分近くに存在し、人工衛星より時刻が含まれた情報を受信するＧＰＳ受信部２１と、受信した情報から時刻を求める機構と、情報を受信できない場合にも自身で継続して時刻を生成できる機構と、ディスクサブシステム３に時刻を配信する時刻配信部２２を有する。

【００２５】一般にホストコンピュータと呼ばれる上位装置１がディスクサブシステム３に対し実行する更新処理の頻度は、各々のシステムによって異なるが、高頻度と言われる場合はマイクロ秒程度の間隔で次々と更新処理が実行されることがある。つまり各ディスクサブシステム３の時計１５をマイクロ秒程度の精度の範囲まで正確な時刻に一致させる理由は、マイクロ秒程度の間隔で次々と実行される更新処理の場合も、その更新処理時刻の差異をディスクサブシステム３が確実に把握できるようにし、更新処理時刻をもとに当該データの更新処理順序を把握できるようにするためである。もしホストコンピュータがディスクサブシステム３に対し実行する更新処理の頻度がマイクロ秒程度の間隔よりもさらに短い時間間隔であれば、各ディスクサブシステム３の保有する時計１５も当該時間間隔以下の時間単位で時刻を一致させる必要がある。

【００２６】図3は、ディスクサブシステム３とディスクサブシステム７に亘ってデータの二重化をする処理の手順を示す図である。なお以下の処理をする前の初期条件として、ディスクサブシステム３とディスクサブシステム７との間でそれ以前のデータ追加・更新の結果が反映され、データの二重化が完了しているものとする。上位装置１がディスクサブシステム３にデータの書込要求（以下、ライトコマンドという）を発行する（ステップ３１）。ディスクサブシステム３がインタフェース制御部１１を介してこのライトコマンドを受領すると、ディスクサブシステム３のディスクサブシステム制御部１７は、そのライトコマンドに基づき処理を開始する。ここでライトコマンドとは、データをキャッシュメモリ１２に書き込むための指示と書き込みデータそのものとを転送するコマンドを指す。ディスクサブシステム３のディスクサブシステム制御部１７は、ライトコマンドを受領した時刻を時計１５から取得し（ステップ３２）、キャッシュメモリ１２にデータを格納し（ステップ３３）、制御メモリ１４にデータの管理情報を格納する（ステップ３４）。この管理情報には、磁気ディスクドライブ１３へのデータ書き込み先のアドレス情報、当該データのライトコマンド受領時刻、ライトコマンド受領通番およびキャッシュメモリ１２上のデータへのポインタが含まれる。ここでライトコマンド受領通番は、当該ディスクサブシステム３が受領したライトコマンドについて付与した一連の番号である。この後、受領したライトコマンドに対する処理の完了を上位装置１に報告する（ステップ３５）。キャッシュメモリ１２に格納したデータは後に磁気ディスクドライブ１３に記録されるが、これは従来の技術であるため、本例では詳細に記述しない。

【００２７】ディスクサブシステム制御部１７は、制御メモリ１４上の制御ビットを参照し、リモートコピーが停止状態でなければ（ステップ３６Ｎｏ）、上位装置１からのライトコマンドに対する処理とは非同期に、インタフェース制御部１１及びインタフェースケーブル４を介して、接続されているディスクサブシステム７に対しライトコマンドを発行する（ステップ３７）。ここでライトコマンドとは、上位装置１から受領したライトコマンドと同様に当該データを書き込む為の指示と書き込みデータそのものとを転送するコマンドを含むが、さらに当該データのデータ受領時刻とデータ受領通番も含まれる。またリモートコピーが停止状態の場合には（ステップ３６Ｙｅｓ）、磁気ディスクドライブ１３へのデータ書き込み先のアドレス情報を自ディスクサブシステム３内の制御メモリ１４に格納し（ステップ３８）、リモートコピーが正常状態に戻った後に相手ディスクサブシステム７にライトコマンドを発行する。

【００２８】ディスクサブシステム７は、インタフェース制御部１１を介してディスクサブシステム３から発行されたライトコマンドを受領する（ステップ４１）。ライトコマンドには、データの他に当該データをメインセンターのディスクサブシステムが受領した時刻と、当該データをメインセンターのディスクサブシステムが受領した通番が含まれている。ディスクサブシステム７は、受領したライトコマンドをもとに処理を開始する。ディスクサブシステム７のディスクサブシステム制御部１７は、受領したライトコマンドをデータ受領通番の順に並べ替えて、通番の抜け、すなわちライトコマンドの抜けがないかどうかチェックする。その後、受領したデータを「仮のデータ」としてキャッシュメモリ１２に格納し、制御メモリ１４にデータの管理情報を格納し、相手ディスクサブシステム３へデータ受領の完了報告を行う（ステップ４２）。

【００２９】次にディスクサブシステム７のディスクサブシステム制御部１７は、受領したライトコマンドに含まれる時刻に基づいてデータの中から正当化時刻を決めるための候補を選定し、他のディスクサブシステム７と協同して正当化時刻候補の裁定を行う（ステップ４３）。正当化時刻候補の裁定については後述する。このようにして正当化時刻が決まったとき、その時刻以前の受領時刻をもつデータを「正式なデータ」としてキャッシュメモリ１２に格納する（ステップ４４）。キャッシュメモリ１２に格納されたデータは後に磁気ディスクドライブ１３に記録される。

【００３０】図4は、受領時刻と受領通番の付加されたデータがリモートコピーされるまでのデータの流れを例によって説明する図である。上位装置１に５つの書き込みデータＤ１，Ｄ２，Ｄ３，Ｄ４及びＤ５がこの順に発

生し、ディスクサブシステム３－１及びディスクサブシステム３－２に順次ライトコマンドを発行したものとする。ディスクサブシステム３－１は、データＤ１，Ｄ３及びＤ５を受領し、順に受領通番と受領時刻、Ｓ１／Ｔ１，Ｓ２／Ｔ３及びＳ３／Ｔ５を付与するものとする。一方ディスクサブシステム３－２は、データＤ２及びＤ４を受領し、順に受領通番と受領時刻、Ｓ１／Ｔ２及びＳ２／Ｔ４を付与する。ディスクサブシステム３－１は、受領したデータをその通番順にキャッシュメモリ１２に格納した後に、各データについてディスクサブシステム７－１にライトコマンドを発行する。一方ディスクサブシステム３－２も受領したデータをその通番順にキャッシュメモリ１２に格納した後に、各データについてディスクサブシステム７－２にライトコマンドを発行する。

【００３１】ディスクサブシステム７－１は、受領通番と受領時刻が付与されたデータを受領し、「仮のデータ」としてキャッシュメモリ１２に格納する。またディスクサブシステム７－２も受領通番と受領時刻が付与されたデータを受領し、「仮のデータ」としてキャッシュメモリ１２に格納する。次にディスクサブシステム７－１が受領したデータの中で最新のデータＤ５とディスクサブシステム７－２が受領した最新のデータＤ４について、ディスクサブシステム７－１，ディスクサブシステム７－２間で両者の時刻を比較する。ここでデータＤ５及びＤ４を正当化時刻候補という。この例ではＤ４に付与された時刻Ｔ４がＤ５に付与された時刻Ｔ５より古いと判定されるので、データＤ４に付与された時刻Ｔ４が正当化時刻として決定され、それ以前の時刻が付与されたデータＤ２及びＤ４が「正式のデータ」としてそのキャッシュメモリ１２に反映される。また時刻Ｔ４以前の時刻が付与されたデータＤ１及びＤ３が「正式のデータ」としてそのキャッシュメモリ１２に反映される。

【００３２】上記の処理の結果としてデータＤ５が「仮のデータ」のまま残ることになり、少なくともデータＤ５はデータＤ４より後にそのキャッシュメモリ１２に反映される。このようにデータＤ４とＤ５の更新の順序性が守られることによって、障害によってリモートコピーが中断したときにデータ回復とリモートコピー再開が可能となる。これに反してデータＤ４より先にデータＤ５が更新し反映された後に障害が生じたとすれば、一般にデータが消失する可能性が生じ、データ回復とリモートコピー再開が困難となる。一方付与された時刻が正当化されたディスクサブシステム７－１のデータＤ１とＤ３及びディスクサブシステム７－２のデータＤ２とＤ４の更新の順序は、ディスクサブシステム３－１のデータＤ１とＤ３及びディスクサブシステム３－２のデータＤ２とＤ４の更新の順序通りとなり、更新の順序性が守られることになる。

【００３３】以上述べたように受領通番は１台のディスクサブシステム３内の更新の順序、従って１台のディスクサブシステム７内の更新の順序を示す番号である。そしてこの受領通番はデータ抜けのチェックのためにも使用される。これに対して受領時刻は複数のディスクサブシステム３に亘る更新の順序を示す数値であり、複数のディスクサブシステム７に亘る正当化時刻の裁定のために使用される。

【００３４】図５は、正当化時刻の裁定に関するディスクサブシステム７の処理の手順を示す図である。正当化時刻候補の裁定要求を発行するディスクサブシステム７のディスクサブシステム制御部１７は、自装置における正当化時刻候補を選定する（ステップ５１）。すなわちデータ受領の通番の順に並んだライトコマンドの中で最新の受領時刻が付与されたライトコマンドを選定する。次にインタフェースケーブル５によって接続された隣のディスクサブシステム７に選定された正当化時刻候補の裁定を要求するコマンドを発行する（ステップ５２）。このコマンドは裁定要求を示すリクエスト、正当化時刻候補に付与された時刻及び当該ディスクサブシステム７の製造装置番号（製番）を含む。またこのとき裁定を要求した時刻を自己の制御メモリ１４内に記憶する。

【００３５】隣のディスクサブシステム７は、このコマンドを受け取り、コマンド中の製番は自装置の製番に一致するか否か判定する（ステップ５３）。自装置の製番でなければ（ステップ５３Ｎｏ）、上記の処理によって自装置における正当化時刻候補を選定する（ステップ５４）。次に受領した正当化時刻候補の時刻と自装置の正当化時刻候補の時刻とを比較判定する（ステップ５５）。自装置の時刻の方が古ければ（ステップ５５Ｙｅｓ）、裁定対象の正当化時刻候補の時刻と製番を自装置の正当化時刻候補の時刻と製番によつて置き換える（ステップ５６）。自装置の時刻の方が新しければ（ステップ５５Ｎｏ）、ステップ５７へ行く。次に自装置の内容に入れ替えしたもの又は受領したそのままの正当化時刻候補の裁定を要求するコマンドをインタフェースケーブル５によって接続された隣のディスクサブシステム７に送信する（ステップ５７）。またこのとき裁定を要求した時刻を自己の制御メモリ１４内に記憶する。ステップ５７の処理の後にステップ５３に戻る。

【００３６】裁定要求コマンドを受け取り、コマンド中の製番が自装置の製番に一致したとき（ステップ５３Ｙｅｓ）、裁定要求した正当化時刻候補を正当化時刻に決定する（ステップ５８）。すなわちディスクサブシステム７の各々の装置で正当化時刻候補が選定されるが、その中で最も古い時刻が正当化時刻として決定される。

【００３７】図６は、正当化時刻の通知に関するディスクサブシステム７の処理の手順を示す図である。正当化時刻を決定したディスクサブシステム７は、正当化時刻以前の時刻が付与されたデータを「正式なデータ」としてキャッシュメモリ１２に格納する（ステップ６１）。

次に隣のディスクサブシステム７に決定した正当化時刻を通知するコマンドを発行する（ステップ６２）。このコマンドは正当化時刻通知を示すリクエスト、決定された正当化時刻及び当該ディスクサブシステム７の製番を含む。またこのとき通知コマンドを発行した時刻を自己の制御メモリ１４内に記憶する。

【００３８】隣のディスクサブシステム７は、このコマンドを受け取り、コマンド中の製番は自装置の製番に一致するか否か判定する（ステップ６３）。自装置の製番でなければ（ステップ６３Ｎｏ）、ディスクサブシステム３から受け取った正当化時刻以前の時刻が付与されたデータを「正式のデータ」としてキャッシュメモリ１２に格納する（ステップ６４）。次に隣のディスクサブシステム７に受領したコマンドをそのまま送信することによって正当化時刻を通知し（ステップ６５）、ステップ６３に戻る。

【００３９】正当化時刻通知のコマンドを受け取り、コマンド中の製番が自装置の製番に一致したとき（ステップ６３Ｙｅｓ）、正当化時刻以前の時刻が付与されたデータについて、不要になった製番等の情報を制御メモリ１４から消去する（ステップ６６）。以上の処理の結果として、複数のディスクサブシステム７に亘って正当化時刻以前の時刻が付与されたライトコマンドはすべて各ディスクサブシステム７内のデータ更新に反映され、正当化時刻より後の新しいライトコマンドが次の正当化時刻候補の選定対象として残されることになる。

【００４０】なお正当化時刻の裁定要求の起動については、最初に起動するディスクサブシステム７を決めておき、以後正当化時刻を決定したディスクサブシステム７が次の正当化時刻の裁定要求を起動する方式がある。あるいは１台のディスクサブシステム７が周期的に正当化時刻の裁定要求を起動してもよい。いずれにしても正当化時刻の裁定と決定した正当化時刻の通知が周期的に行われるような方式であればよい。

【００４１】また上記実施形態では各ディスクサブシステム７を判別するための識別子として製造装置番号を利用するが、製造装置番号でなくともリモートセンターの各々のディスクサブシステム７を判別することができる識別子であれば、その識別子を利用することができる。

【００４２】正当化時刻の裁定要求を発行してから所定時間内に自己の発行した正当化時刻裁定要求コマンドが戻らないか又は正当化時刻通知を受け取らない場合、あるいは正当化時刻通知を発行してから所定時間内に自己の発行した正当化時刻通知が戻らない場合には、ディスクサブシステム７は、何らかの障害発生により時刻の裁定又は時刻の通知が完了しなかったものと判断し、制御メモリ１４の制御ビットをリモートコピー停止状態に設定し、リモートコピーを停止させる。このようにディスクサブシステム７側の障害によりリモートコピーを続行できなくなったとき、リモートコピーを停止させ、相手ディスクサブシステム３にリモートコピーの停止を通知する。これを受け取ったディスクサブシステム３は、制御メモリ１４上の制御ビットをリモートコピー停止状態に設定し、リモートコピーを停止させ、上記のように上位装置１から受領したライトコマンドのディスクサブシステム７への転送を保留する。ディスクサブシステム３側の障害によりリモートコピーを続行できなくなったときも、同様にリモートコピーを停止させ、相手ディスクサブシステム７にリモートコピーの停止を通知する。あるいはインタフェースケーブル４の障害によりディスクサブシステム３とディスクサブシステム７間の通信不可を検出したディスクサブシステム３及びディスクサブシステム７も同様にリモートコピーを停止させる。また任意の時点でディスクサブシステム３又はディスクサブシステム７のサービスプロセッサパネル１８からの指示によって制御メモリ１４上の制御ビットを変更し、リモートコピー停止状態に設定したり、同停止状態を解除してリモートコピーを再開させることができる。

【００４３】なおリモートコピーの一時停止や再開は、リモートコピーを設定する際のディスクサブシステム内の記憶領域の最小管理単位であるボリュームペア単位に指定できる。ここでボリュームペアとは、コピー元のディスクサブシステム３内ボリュームと対応するコピー先のディスクサブシステム７内ボリュームとのペアをいう。少なくとも１つのボリュームペアを一つのボリュームグループとして定義し、ボリュームグループ単位に状態を変化させることも可能である。ここでボリュームグループ内のデータ更新は他のボリュームグループに影響しないものとしている。この場合には、制御メモリ１４上の制御ビットはボリュームペアあるいはボリュームグループごとに設ける。またディスクサブシステム３は、ステップ３６でデータを格納すべきボリュームあるいはボリュームグループと制御メモリ１４の対応する制御ビットを参照して、当該データのリモートコピーが停止状態か否かを判定する。

【００４４】従って例えば複数の業務データが格納されているディスクサブシステムの全データがリモートコピーで常時二重化されているシステムにおいて、ある業務については業務終了時点の状態のデータをリモートセンター側で利用したい等の用途がある場合、業務終了時点で当該業務データが格納されているボリュームペア又はボリュームグループについてリモートコピーを一時停止状態にすれば、リモートセンターのディスクサブシステム７に業務終了時点のデータの状態が保持されるため、リモートセンターで業務を行うことができる。

【００４５】一時停止状態が解除されると、ディスクサブシステム３及び７は、リモートコピー処理を再開するとともに、一時停止している間にメインセンターのディスクサブシステム３のみで書き込みが行われたデータやリモートセンターのディスクサブシステム７で正当化が

完了していなかったデータについて、ライトコマンドをディスクサブシステム７に転送したり、データの正当化処理を再開する。これにより通常のリモートコピーが行われる状態に復帰できる。

【００４６】上記実施形態は、メインセンター９側のディスクサブシステム３とリモートセンター１０側のディスクサブシステム７が１対１に対応して接続されるシステム構成についての実施形態であるが、他の実施形態として、複数のディスクサブシステム３が１台のディスクサブシステム７に接続され、このディスクサブシステム７が接続される複数のディスクサブシステム３に関するリモートコピー処理をするように構成してもよい。この構成において、ディスクサブシステム７は、ステップ４１のデータ受領のとき、受領したライトコマンドを各ディスクサブシステム３の製造装置番号ごとにデータ受領通番の順に並べ替え、各ディスクサブシステム３のライトコマンドに付与されている時刻のうち最新の時刻を選定し、さらにディスクサブシステム３間で選定された時刻を比較してその中から最古の時刻を正当化時刻候補として選定し、上記の正当化時刻の裁定処理を行う。

【００４７】また他の実施形態として、リモートセンターのディスクサブシステム７のいずれか１台をマスタディスクサブシステム（一般にマスタ記憶サブシステム）に設定し、リモートセンター１０のマスタディスクサブシステム以外の各ディスクサブシステム７とマスタディスクサブシステムとを互いに接続するような構成をとってもよい。この構成においてマスタディスクサブシステムは、リモートセンター１０の他のディスクサブシステムが自身で保有する正当化時刻候補を問い合わによって取得する。次にマスタディスクサブシステムは、取得した正当化時刻候補及び自己の正当化時刻候補を比較して最古の時刻を正当化時刻として決定する裁定を行い、リモートセンタの他のディスクサブシステムに決定した正当化時刻を通知し、各ディスクサブシステム７が各々データの正当化を実施する。

【００４８】本実施形態によれば、リモートセンター１０のディスクサブシステム７が正当化時刻候補の裁定および正当化時刻に基づくデータの正当化を実施することにより、上位装置の介在なくまたリモートコピーを定期的に中断させる処理を行うことなく、さらにメインセンター９とリモートセンター１０の各々に更新順序を保持するためのゲートウェイサブシステム等を設けることもなく、メインセンター９のディスクサブシステム３の処理性能の低下なく、ディスクサブシステムの機能で更新順序の一貫性を保持した非同期型のリモートコピーを実現できる。

【００４９】また本実施形態によれば、ディスクサブシステム３が保有する時計から取得した時刻を利用して更新データの一貫性を保証するため、上位装置１からディスクサブシステム３へ転送されるデータに時刻情報が付加されないシステムであっても、上位装置１の介在なしに更新順序の一貫性を保持した非同期型のリモートコピーを実現できる。

【００５０】また仮にメインセンターの被災や機器の障害等によりメインセンター９の機能が絶たれた場合も、リモートセンター１０のディスクサブシステム７ではデータの一貫性が保たれた状態のデータが記録されている。これらはすべてディスクサブシステム３，７の機能のみで実現され、上位装置１の処理能力に対し負担とならない。メインセンター９が被災した場合は、ディスクサブシステム７のデータを利用してジョブを再実行する等の回復作業を行い、業務を再開できる。

【００５１】

【発明の効果】以上述べたように本発明によれば、上位装置に新規ソフトウェアの導入を必要とせず、記憶サブシステムの機能のみでユーザが期待する範囲での更新データの一貫性を保証でき、導入が容易でかつメインセンターのホストコンピュータの処理性能の低下がない非同期型のリモートコピーシステムを実現できる。

【図面の簡単な説明】

【図１】実施形態のリモートコピーシステムの全体構成を示す図である。

【図２】実施形態のディスクサブシステムの内部構成を示す図である。

【図３】実施形態のディスクサブシステム３とディスクサブシステム７に亘ってデータの二重化をする処理の手順を示す図である。

【図４】受領時刻と受領通番の付加されたデータがリモートコピーされるまでのデータの流れを例によって説明する図である。

【図５】実施形態の正当化時刻の裁定に関するディスクサブシステム７の処理の手順を示す図である。

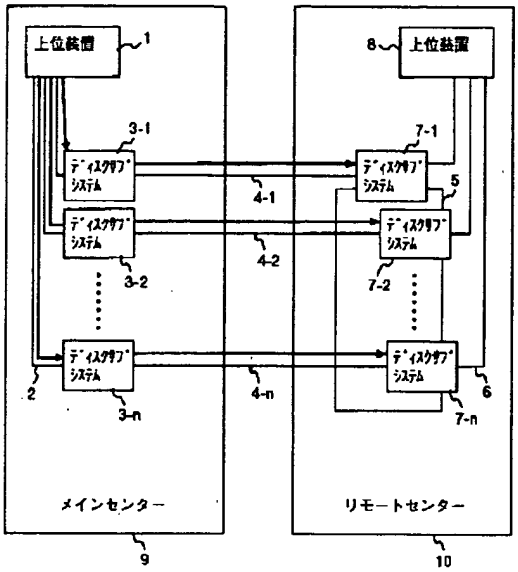
【図６】実施形態の正当化時刻の通知に関するディスクサブシステム７の処理の手順を示す図である。

【符号の説明】

１…上位装置、３…ディスクサブシステム、４…インタフェースケーブル、５…インタフェースケーブル、７…ディスクサブシステム、９…メインセンター、１０…リモートセンター、１５…時計、１６…時刻補正部

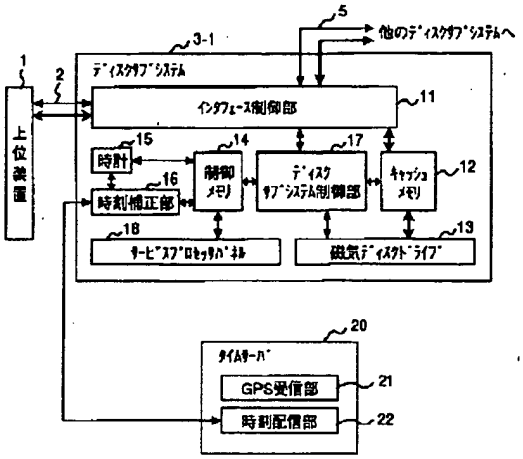
【図1】

図 1



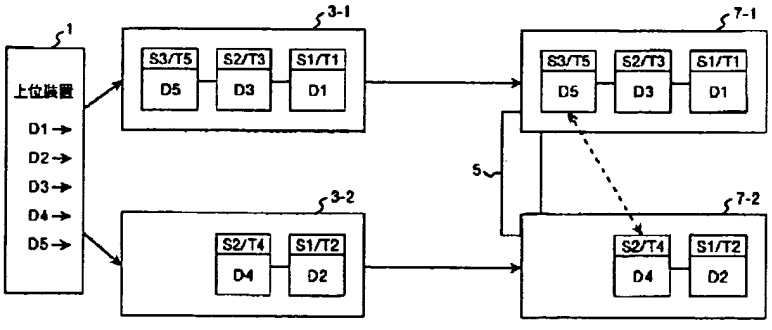

【図2】

図 2



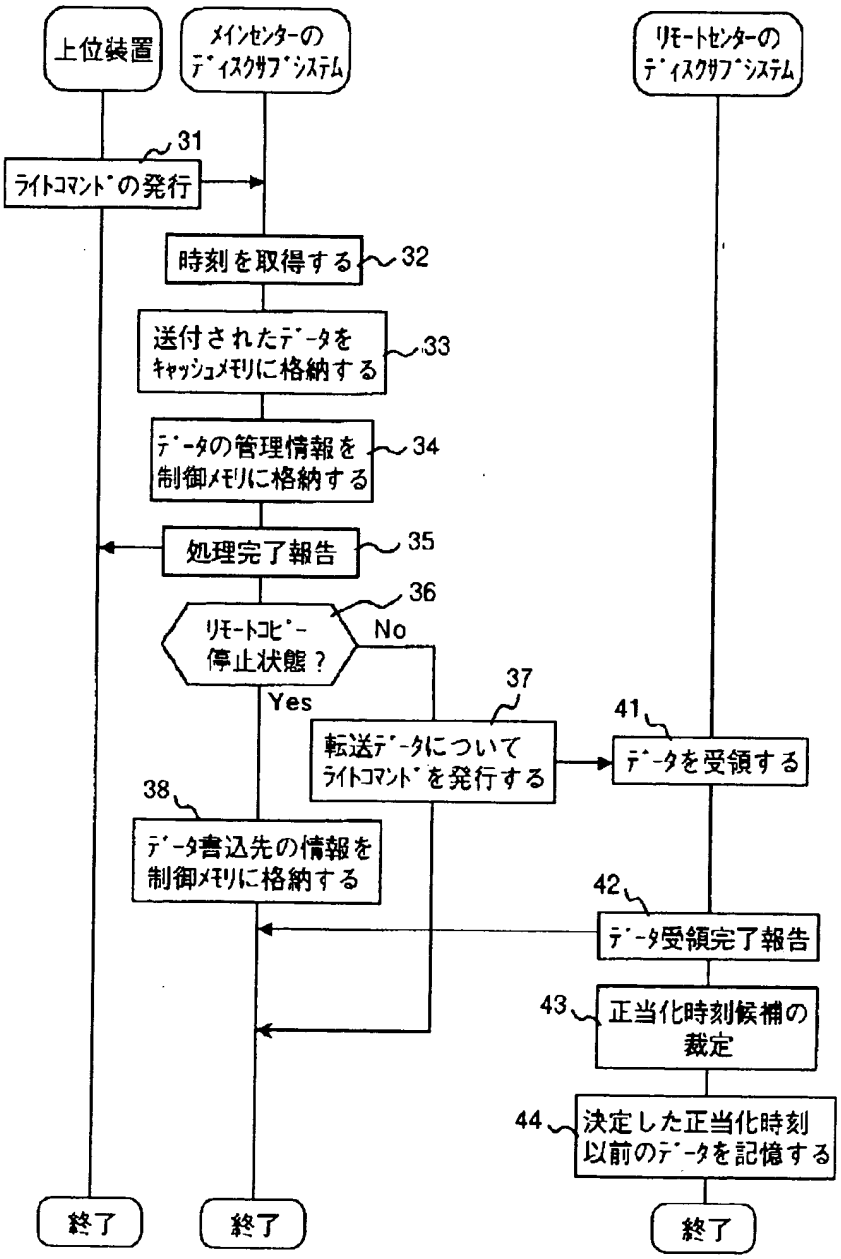

【図4】

図 4



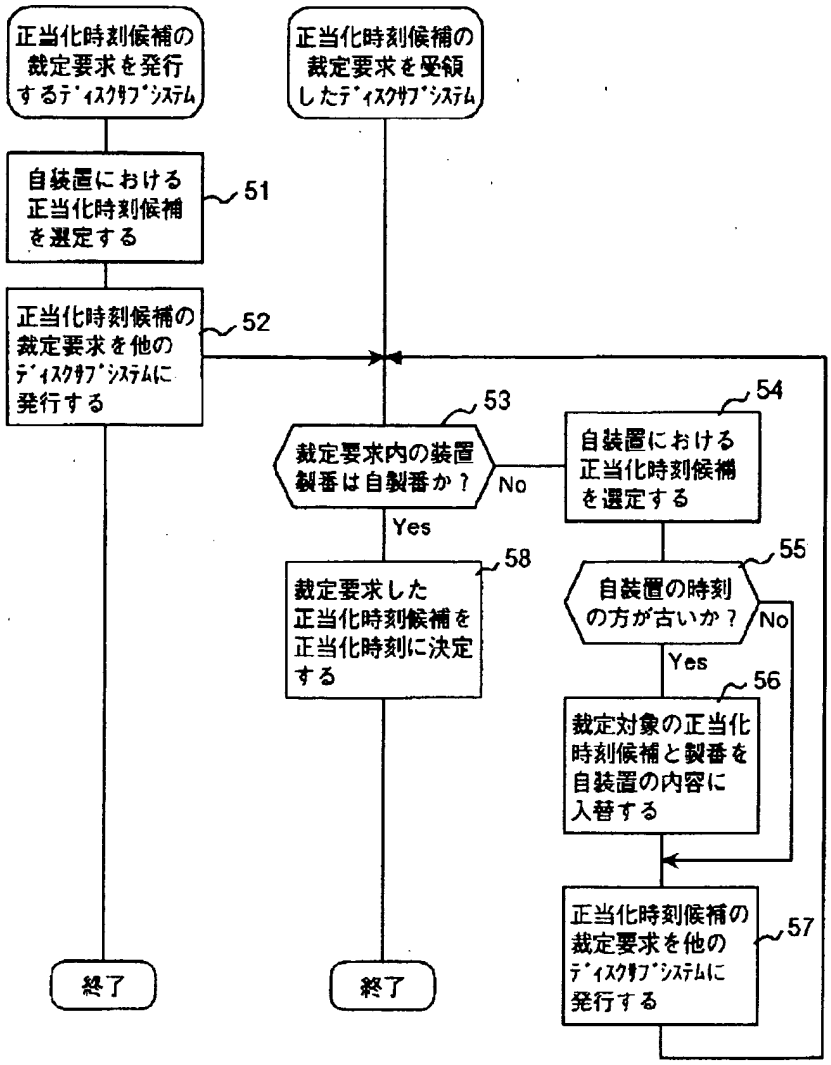
【図3】

図 3


上位装置
メインセンターのディスクサブシステム
リモートセンターのディスクサブシステム
31
ライトコマンドの発行
時刻を取得する
32
送付されたデータをキャッシュメモリに格納する
33
データの管理情報を制御メモリに格納する
34
処理完了報告
35
36
リモートコピー停止状態？
No
Yes
37
転送データについてライトコマンドを発行する
41
データを受領する
38
データ書込先の情報を制御メモリに格納する
42
データ受領完了報告
43
正当化時刻候補の裁定
44
決定した正当化時刻以前のデータを記憶する
終了
終了
終了

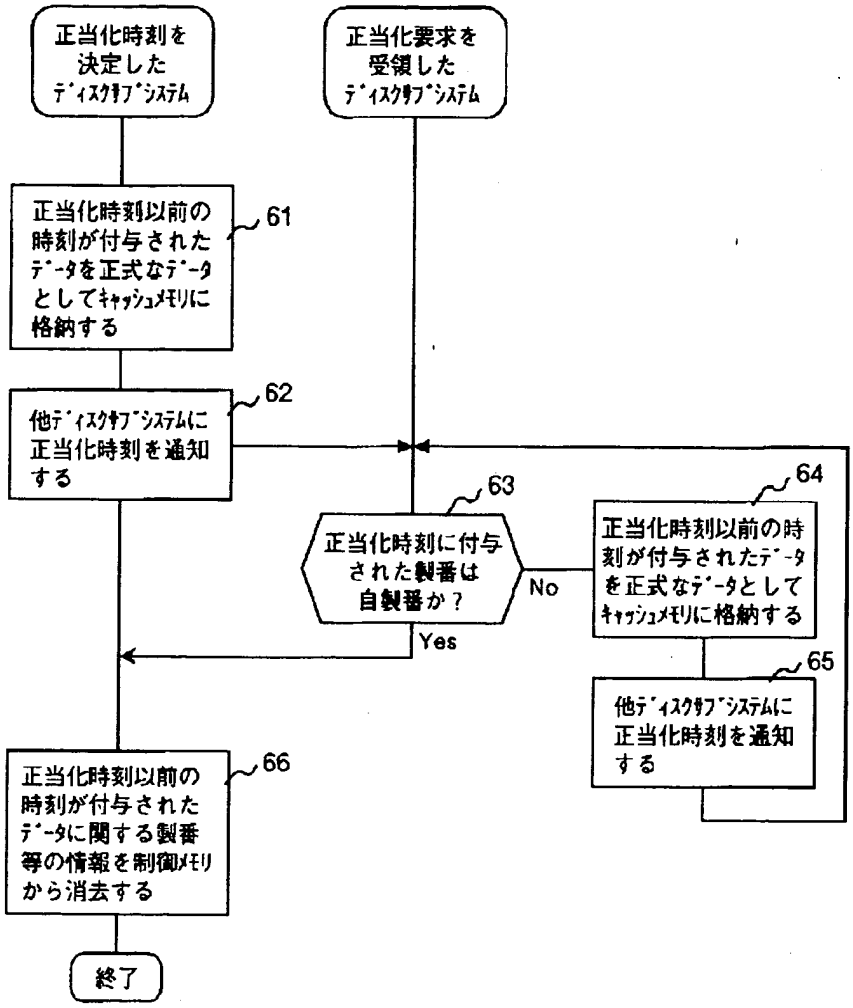
【図5】

図 5


正当化時刻候補の裁定要求を発行するﾃﾞｨｽｸｻﾌﾞｼｽﾃﾑ
正当化時刻候補の裁定要求を受領したﾃﾞｨｽｸｻﾌﾞｼｽﾃﾑ
自装置における正当化時刻候補を選定する
51
正当化時刻候補の裁定要求を他のﾃﾞｨｽｸｻﾌﾞｼｽﾃﾑに発行する
52
53
裁定要求内の装置製番は自製番か？
No
Yes
54
自装置における正当化時刻候補を選定する
55
自装置の時刻の方が古いか？
No
Yes
56
裁定対象の正当化時刻候補と製番を自装置の内容に入替する
58
裁定要求した正当化時刻候補を正当化時刻に決定する
正当化時刻候補の裁定要求を他のﾃﾞｨｽｸｻﾌﾞｼｽﾃﾑに発行する
57
終了
終了

【図6】

図 6




---

フロントページの続き

(72)発明者 中野 俊夫
神奈川県小田原市国府津2880番地 株式会社日立製作所ストレージシステム事業部内

Fターム(参考) 5B018 GA06 HA03 HA04 HA31 JA26 KA02 KA22 MA12 PA03 RA11
5B034 BB17 CC01 CC02
5B065 BA01 CE22 EA02 EA12 EA35
5B082 DE04 GB02
5B083 AA08 AA09 BB03 CC04 CD11 EE08

Japanese Published Unexamined Patent Application (A) No. 13-282628; published October 13, 2001; Application Filing No. ,112-101168, filed March 31, 2000; Inventor(s): Hideo Tabuchi et al.; Assignee: Hitachi Manufacturing, Inc.; Title of Invention: Data Duplication System and Method in Memory Subsystems

---

# DATA DUPLICATION SYSTEM AND METHOD BY MEMORY SUBSYSTEMS

## CLAIM(S)

1) A data duplication method using a first memory subsystem group of multiple memory subsystems and a second memory subsystem group of multiple memory subsystem, in which to store the copies of data of the first memory subsystem group, comprising the following steps:

by each memory subsystem in the first memory subsystem group, multiple data, each attached with a serial number and time, are input in each memory device of each memory subsystem and transferred to the memory subsystems in the second memory subsystem group via a transmission path;

multiple data received by each memory subsystem in the second memory subsystem group are arranged in a sequence of serial numbers;

through communication among the memory subsystems in the second memory subsystem group, the oldest time is selected out of the newest times provided by said memory subsystems;

the data attached with the time preceding the selected oldest time is input in each memory device of each memory subsystem.

2) A data duplication method using memory subsystems, as cited in Claim 1, wherein said transmission path for connecting said first memory subsystem group and said second memory subsystem group constitutes a storage area network (SAN).

3) A data duplication method using memory subsystems, as cited in Claim 1, wherein a time to be referred to by each memory subsystem in said first memory subsystem group is corrected by external time data.

4) A data duplication method using memory subsystems, as cited in Claim 1, wherein the memory subsystems in said second memory subsystem group are connected with a loop transmission path; each memory subsystem transfers to the adjacent memory subsystem the older time out of its own newest time and said received newest time, and the time that is transmitted and returned to its own memory subsystem is determined as said oldest time.

5) A data duplication method using memory subsystems, as cited in Claim 1, wherein one of multiple memory subsystems in said second memory subsystem group is selected as a master memory subsystem; said newest time is notified to the master memory subsystem by each memory subsystem other than the master memory subsystem; the master subsystem

determines the oldest time out of its own newest time and the acquired newest time.

6) A data duplication method using memory subsystems, as cited in Claim 1, wherein every data input in every memory subsystem in the first subsystem group is transferred, by every memory subsystem in said first memory subsystem in said first memory subsystem group, to one memory subsystem in said second memory subsystem group; the newest time is selected out of the times provided by all the memory subsystems in said first memory subsystem group by one memory subsystem in the second memory subsystem group, and the oldest time selected out of the newest times is determined as a candidate for the oldest time.

7) A data duplication method using memory subsystems, as cited in Claim 1, the memory device of said memory subsystem consists of multiple volumes, and when a volume in the first memory subsystem group to be copied from and a volume in the second memory subsystem to copy to constitute a volume pair, starting and stopping of the data transfer from the memory subsystems in the first memory subsystem group to the second memory subsystem group are controlled per a volume group consisting of multiple volume pairs.

8) A data duplication system using memory subsystems, comprising a first memory subsystem group consisting of multiple memory subsystem and a second memory subsystem group consisting of multiple memory subsystems, wherein each memory subsystem in the first memory subsystem group has a means of inputting data in its memory device, a means of attaching a serial number and a time to the data, and a means of transferring the data attached with the serial number and time to the memory subsystem in the second memory subsystem via the transmission path; each memory subsystem in the second memory subsystem group has a means of arranging the received multiple data in a sequence of serial numbers, a means of determining the oldest time out of the newest times provided by the memory subsystems by communication among the memory subsystems in the second memory subsystem group, and a means of inputting the data attached with the time preceding the determined oldest time in each memory device of each memory subsystem.

9) A data duplication system using memory subsystems, as cited in Claim 8, wherein said transmission path for connecting the memory subsystems in said first memory subsystem group and the memory subsystems in said second memory subsystem group constitutes a storage area network (SAN).

10) A data duplication system using memory subsystems, as cited in Claim 8, wherein each memory subsystem in said first memory subsystem group has a means of correcting said reference time for said times based on the externally supplied time data.

11) A data duplication system using memory subsystems, as cited in Claim 8, wherein the memory subsystems in said second memory subsystem group are connected with a loop transmission path; each memory subsystem has a means of transferring to the adjacent memory subsystem the older time out of its own newest time and said received newest time and of determining the time transmitted and returned as the oldest time.

12) A data duplication system using memory subsystems, as cited in Claim 8, wherein one of multiple memory subsystems in said second memory subsystem group is selected as a master memory subsystem; each memory subsystem other then said master memory subsystem has a means of notifying the master memory subsystem of said newest time; said master memory subsystem has a means of determining said oldest time out of its own newest time and acquired newest time.

13) A data duplication system using memory subsystems, as cited in Claim 8, wherein each of multiple memory subsystems in said first memory subsystem group is structured to transfer each of the input data to one

memory subsystem in said second memory; said one memory subsystem in the second memory subsystem group has a means of selecting the newest time out of the times provided by said memory subsystems in the first memory subsystem group and a means of determining the oldest time out of the selected newest times as the oldest time candidate.

14) A data duplication system by memory subsystems, as cited in Claim 8, wherein the memory device of said memory subsystem consists of multiple volumes; when a volume in the first memory subsystem group to be copied from and a volume in the second memory subsystem to copy to constitute a volume pair, the memory subsystems in the first memory subsystem group has a means of controlling the starting and stopping of data transfer to the second memory subsystem group per a volume group consisting of multiple volume pairs.

15) A memory subsystem in a memory subsystem group consisting of multiple memory subsystems constituting a data duplication system, wherein said memory subsystem has a first means of inputting its externally received data in its memory device, a second means of attaching a serial number and time to said data and of transferring it to other memory subsystem, and a third means of arranging multiple data received from other memory subsystem in a sequence of serial numbers, and a fourth means of

determining the oldest time out of the newest times provided by other memory subsystems through communication among them; when said memory subsystem is in a local mode, it activates the first means and the second means, and in a remote mode, it activates the third means, the fourth means, and the first means; it inputs by the first means the data attached with the time preceding the determined oldest time.

16) A memory subsystem constituting a data duplication system, as cited in Claim 15, wherein said memory subsystem operating in said local mode and said memory subsystem operating in said remote mode are connected with a storage area network (SAN).

17) A memory subsystem constituting a data duplication system, as cited in Claim 15, wherein said memory subsystem has a means of correcting said reference time based on the externally supplied time data.

18) A memory subsystem constituting a data duplication system, as cited in Claim 15, wherein said memory subsystems operating in said remote mode are connected with a loop transmission path, and each memory subsystem in said remote mode has a means of transferring the older time out of its own newest time and the received newest time to its adjacent memory subsystem and a means of determining the time transmitted and returned to its own memory subsystem as said oldest time.

19) A memory subsystem constituting a data duplication system, as cited in Claim 15, wherein one of said memory subsystems operating in said remote mode is selected as a master memory subsystem; each memory subsystem other than said master subsystem has a means of notifying said newest time to the master memory subsystem, and the master memory subsystem has a means of determining said oldest time out of its own newest time and the acquired newest time.

20) A memory subsystem constituting a data duplication system, as cited in Claim 15, wherein multiple memory subsystems operating in said local mode are structured to transfer the data to one memory subsystem in the memory subsystem group operating in said remote mode; said one memory subsystem in the memory subsystem group operating in said remote mode has a means of selecting the newest time out of the times provided by the memory subsystems operating in said local mode and a means of determining the oldest time out of the selected newest times as said oldest time candidate.

21) A memory subsystem constituting a data duplication system, as cited in Claim 15, wherein the memory device of said memory subsystem consists of multiple volumes; when a volume to be copied from, which is in the memory subsystem and operates in said local mode, and a volume to be

copied to, which is in the memory subsystem and operates in said remote mode, constitute a volume pair, the memory subsystem operating in said local mode has a means of starting and stopping the data transfer to the memory subsystem operating in said remote mode per every volume group consisting of at least one volume pair.

## DETAILED DESCRIPTION OF THE INVENTION

(0001)

(Field of Industrial Application)

The present invention pertains to a memory subsystem for a computer to refer to and update data in, particularly to a data duplication system and its method for duplicating the data stored in the memory subsystem.

(0002)

(Prior Art)

It is already known that there are companies and monetary institutions that have computer centers and back-up computer centers equipped with systems for backing-up data against earthquakes and disasters. In such a data back-up system, a primary computer center and a remote computer center are set up geographically apart by at least a distance of one hundreds kilometers - a few hundreds kilometers. Between disk subsystems of the primary center and of the remote center, data are stored and duplicated. In

one of its methods, data written in the disk subsystem of the primary center are transferred to the disk subsystem of the remote center in order to write the same data in the disk subsystem of the remote center. The systems having such a remote copying function have already been put to practical use. A remote copying function is primarily divided into two types: a synchronous type and an asynchronous type.

(0003)

A synchronous type refers to a process, wherein when a host computer (higher device) in the primary center commands its disk subsystem to update data (to input data) and when the object to which the command is given is a remote-copying function, the remote center, after having completed the updating and the remote copying of said data, which have been commanded to the disk subsystem of the remote center, reports the completion of the data-updating process to the higher device of the primary center. In such a case, a physical distance between the primary center and the remote center has an impact on the performance of a data transmission path, generating a delay in transmission time.

(0004)

An asynchronous type refers to a process, wherein when the higher device in the primary center commands its disk subsystem to update data (to

input data), completion of the data updating is reported to the higher device as soon as the data has been updated even if the object to which the command is given is a remote-copying function, and the data updating (copying) in the disk subsystem at the remote center is executed asynchronously with the process at the primary center. By the asynchronous process, the data updating is completed within the processing time required in the primary center, so the time delay that is caused by storing the data in the remote center is not generated. Accordingly, when remote copying is performed between the subsystems at the remote sites, an asynchronous remote copying method will be more suitable than a synchronous method if high priority needs to be placed on avoiding an impact of the transmission time on the operation of the primary center.

(0005)

With the asynchronous type, the content of the subsystem in the remote center is not always consistent with that in the primary center. Therefore, if the primary center loses its function due to disasters, the data which are yet to be copied to the remote center will be lost. However, the access performance of the disk subsystem of the primary center can be improved to a level high enough not to use the remote-copying function.

(0006)

The aforementioned prior art, however, comes with various problems since the remote-copying function is accomplished by an intermediate function of a higher device.

(0007)

"Preserving Data Consistency"

When the remote copying is performed, the disk subsystem of the primary center and the disk subsystem of the remote center are connected by an independent communication link. More specifically, when remote copying is performed between multiple subsystems of the primary center and multiple subsystems of the remote center, multiple sets of subsystems, each set being connected with an independent communication link, are used. When the remote center backs up the primary center having multiple subsystems, there is a need of "preserving consistency of data," that is, a need of preserving a consistent sequence of data updating among multiple data subsystems. In asynchronous remote copying, it is unavoidable that updated data is copied to the remote center with a delay after the data updating has actually completed in the primary center. But, a sequence of updating must be consistent with that in the primary center. Even if the sequence of updating data is not always consistent, consistent data must be

at least recorded in the remote center when the user needs the data at the remote center.

(0008)

A data base generally consists of database body, various logging data indicating the history of data updating, and of control data, each being related to one another. The system is often designed so that the database body, various logging data, and control data are recorded in different subsystems, taking reliability into consideration. However, since the database body, various logging data, and control data are related to one another, the logging data and control data need to be added and updated when data is updated, while keeping their coordination in the system. The time interval between one updating and the subsequent updating in a sequence of data updating can be a few micron seconds if the updating is done in a short period of time in succession. If the sequence of updating is disrupted, the coordination of data to preserve the sequence of data updating is also disrupted, possibly destroying the entire database in the worst case.

(0009)

For example, it is possible that, even if the logging data is updated after the database body has been updated at the primary center, the updated logging data may arrive at the remote center before the logging data of the

database body does, depending upon the condition of said communication link constituting the remote copying system; then, the logging data will be added and updated before the database body is added and updated. If a disaster falls on the primary center under such a logically uncoordinated condition wherein only the logging data is added and updated but the database body is not added and updated, the data in the remote center will be useless. Therefore, data must be updated in the remote center exactly in the same sequence as the sequence in which data has been updated in the primary center.

(0010)

"Remote-Copying Function by Insertion of Higher Device"

When asynchronous remote copying is performed in a general environment in which there are multiple disk subsystems in the primary center and in the remote center, there is a method wherein the data regarding the sequence of updating, such as a stamped time, is attached to data when the higher device of the primary center commands the disk subsystems to update data and the higher device of the remote center executes the data updating and copying based on said time data. For example, this technology is disclosed in Japanese Published Unexamined Patent Application 06-290125 (U.S. Patent 5446871). In Japanese Published Unexamined Patent

Application 06-290125, a remote copying operation is performed by inserting the higher device. More specifically, the cooperation between the operating system and subsystems of the primary center and the data mover software and subsystem of the higher device of the remote center implements the issuing and sending of the data regarding the sequence of updated data and the copying of the updated data by use of said data.

(0011)

(Problems of the Prior Art to Be Addressed)

According to the prior art technology, asynchronous remote copying is executed while securing a sequence of updated data between the primary center and the remote center. In this method, however, the higher device and the disk subsystem both have to have a software application to execute the asynchronous remote copying function, and both sides have to cooperate in the operation. Then, it becomes necessary for the user to reconsider the system design because of a need of introduction, installation, and inspection of the software application and of an increase in load on the CPU. Therefore, there is a problem that it takes long time and high cost for introducing the prior art technology. In addition, with the system wherein the higher device does not attach the data regarding the sequence of updating, such as a stamped time, or with the system not equipped with a mechanism

of matching the internal clocks in host computers, and particularly with an open system, there is no means of implementing said prior art technology. Therefore, in order to implement an asynchronous remote-copying function only by using the disk subsystem function, it is necessary to preserve, regardless of types of higher device and types of software application, the consistency in sequence of data updating only by using the disk subsystem function . Also, with the prior art technology, there was a problem that there was no means of preserving the consistency in sequence of updating among multiple disk subsystems when data requiring the consistency in sequence of updating are distributed to and stored in multiple subsystems.

(0012)

The objective of the present invention is to present an asynchronous remote-copying function that can secure consistency of data only by using the function of the subsystem without installing a new software application on the higher device and that can be introduced without reducing the performance of the primary center.

(0013)

(Mean to Solve the Problems)

The present invention presents a data-duplicating method using a system comprising a first memory subsystem group consisting of multiple memory

subsystems and a second memory subsystem group consisting of multiple memory subsystems for storing the copy of data of the first memory subsystem group, said method being characterized by the following steps: data are input in the memory device of each memory subsystem of the first memory subsystem group while simultaneously transferring, via transmission path, said data to the memory subsystem of the second memory subsystem group by attaching a serial number and a time to said data; multiple data received by each memory subsystem of the second memory subsystem group are arranged in a sequence of serial numbers; every memory subsystem of the second memory subsystem group, by communicating with one another, determines the oldest time data out of the time data attached to the data and inputs the data attached with the time preceding said determined oldest time in the memory device of each memory subsystem. The present invention also presents a remote-copying system that functions according to the aforementioned method.

(0014)

(Embodiment Example)

The embodiment example of the system of the present invention is explained below with reference to the drawings.

(0015)

Fig. 1 shows the system structure wherein data is duplicated between two centers each having a computer. Multiple disk subsystems, 3 – 1, 3 – 2, ..., 3 – n, of primary center and the multiple disk subsystems, 7 – 1, 7 – 2, ..., 7 – n, of remote center 10 are connected without using the higher devices 1 and 8, and both centers constitute the remote-copying system for duplicating data. The method of connecting the disk subsystems without using the higher devices 1 and 8 is not herein described in detail but, for example, a storage area network (SAN) can be used for the connection.

(0016)

In the primary center 9, the higher device 1 having a central processing unit (CPU) for data processing is connected to the disk subsystems, 3 – 1, 3 – 2, ..., 3 – n, via the interface cable 2.

(0017)

The disk subsystem 3 – 1 is connected, via the interface cable 4 – 1, to the disk subsystem 7 – 1 of the remote center. The disk subsystem 3 – 2 is connected, via the interface cable 4 – 2, to the disk subsystem 7 – 2 of the remote center. Likewise, the disk subsystem 3 – n is connected, via the interface cable 4 – n, to the disk subsystem 7 – n of the remote center. Hereinafter, one of the disk subsystems, 3 – 1, 3 – 2, ..., 3 – n, or one of the

disk subsystems, 7 – 1, 7 – 2, ..., 7 – n, may be referred to as disk subsystem 3 or disk subsystem 7. The same applies to other components as well.

(0018)

Also, the interface cables, 4 – 1, 4 – 2, ..., 4 – n, include an optic fiber link driven by the LED driving device and a fiber channel driven by an interface protocol by using an optic fiber cable. They may also include a telecommunication link represented by a T3 network or ATM network. Accordingly, a general fiber channel switch may be connected to some point along the interface cable between the disk subsystem 3 of the primary center 9 and the disk subsystem 7of the remote center 10, or the interface cable may be connected to a T3 network, SONET network, and ATM network. In this example, the communication link is referred to as the interface cable 4 provided that it can have any of these connection modes.

(0019)

In the remote center 10, one of the disk subsystems, 7 – 1, 7 –2, 7 – n, to which the data is to be remote copied is, connected, via the interface cable 5, to one of the other subsystems in said center to which data is to be remote copied. In this example, the disk subsystem 7 – 1 is connected to disk subsystem 7 – 2, the disk subsystem 7 – 2 to disk subsystem 7 – 3, and disk

subsystem 7 – n to the disk subsystem 7 – 1, forming a connected continuous loop.

(0020)

The higher device 8 is connected to the disk subsystems, 7 – 1, 7- 2, ..., 7 – n, via the interface cable 6 in the remote center, and it is a central processing unit that looks up and updates data in the disk subsystems, 7 – 1, 7 – 2, ..., 7 – n. The higher device 8 can process data on behalf of the higher device 1 of the primary center 9 when the higher device 1 of the primary center 9 cannot perform its function due to failure or disaster. It also can executes a process different from that of the higher device 1of the primary center 9 independently from the higher device 1. However, if the higher device 8 needs not process data in the disk subsystems, 7 – 1, 7 – 2, ..., 7 – n, the higher device 8 is not necessary.

(0021)

Fig. 1 shows the system structure, wherein, in the case when the higher device 1 commands multiple disk subsystems, 3 – 1, 3 – 2, ..., 3 – n, to input data, the same data are also input in multiple disk subsystems, 7 – 1, 7 – 2, ..., 7 – n of the remote center 10, while preserving consistency in data updating with the updating in the primary center 9. The arrow in Fig. 1

shows a flow of data, which the higher device has commanded to be input in the disk subsystems.

(0022)

Fig. 2 shows the internal structure of the disk subsystem 3 and of disk subsystem 7. The disk subsystem 3 comprises: interface controller 11 for receiving/sending a command and the data transferred from the higher device 1 and for connecting itself to other disk subsystem 7; cache memory 12 for storing the data which is referred to or updated by the higher device 1; magnetic disk drive 13 as a recording medium for recording said data, control memory 14 for storing the control data for said data; the disk subsystem controller 17 for controlling each of these components. The disk subsystem controller 17 is operated by the program control that uses a microprocessor.

(0023)

The disk subsystem 7 has the internal structure similar to that of disk subsystem 3. Also the disk subsystem 3 and the disk subsystem 7 can be structured to have the same internal structure and the same function, and the disk subsystem functions as the disk subsystem 3 in local mode setting and functions as the disk subsystem 7 in remote setting. The functions in local mode and remote mode are explained below later. In addition, a memory

device other than the magnetic disk may take place of general magnetic disk derive 13. For the disk subsystems 3 and 7, the ones generally called memory subsystems may be used.

(0024)

Each disk subsystem 3 is equipped with clock 15 and clock correction section 16. The clock correction section 16 frequently corrects the time of clock 15 based on the time data supplied by the time server located at a close but separate position from each subsystem to reduce the time difference between the time supplied from the time server and the time of clock 15 of each disk subsystem 3 to nearly a few micron seconds. In this context, the timer server has a function to receive radio wave containing the time data like a GPS or microwave clock does and a function to deliver the received time data to each disk subsystem 3. Or it can be the one wherein the time server has a function to generate and deliver the time to each disk subsystem without having the radio wave-reception function. If one of the disk subsystems 3 has the time server function, an independent time server is not needed. The time server 20 shown in Fig. 2 is located at a position close enough to the disk subsystem 3 and has GPS receiving section 21 for receiving data containing the time data from a satellite, a mechanism to seek the time from the received data, a mechanism to continue to generate the

time in the case when the time data cannot be received, and time delivery section 22 for delivering the time to disk subsystems 3.

(0025)

The frequency at which the higher device 1, which is generally referred to as a host computer, executes data updating in the disk subsystem 3 varies depending upon the system, but in the case when this updating is processed very frequently, the updating is performed successively at intervals of nearly micro seconds. The reason for accurately setting the clock 15 of every subsystem as accurately as a micro second range is to help the disk subsystem 3 accurately catch the time difference between the times attached to data even when the updating processes are executed successively at intervals of a few micro seconds and to catch the sequence of updating said data by using the times when the data are updated. If the host computer commands the disk subsystem 3 to execute the updating at closer intervals than micro seconds, the clocks 15 of all the disk subsystems also must match the time among them by using a time unit smaller than said time interval.

(0026)

Fig. 3 shows the steps of process for duplicating data by the disk subsystem 3 and the disk subsystem 7. The initial condition prior to the following process presupposes that the previously added and updated data

have already been copied between the disk subsystem 3 and disk subsystem 7, therefore, the duplication of said data has already been completed. First, the higher device 1 commands the disk subsystem 3 to write data (hereinafter referred to as a write command) (Step 31). When the disk subsystem 3 receives this write command via the interface controller 11, the disk subsystem controller 17 of disk subsystem 3 begins to process said write command. The write command in this context includes the command to write data in cache memory 12 and the command to transfer the data to be written. The disk subsystem controller 17 of the disk subsystem 3 acquires the time when the write command is received from clock 15 (Step 32), stores the data in cache memory 12 (Step 33), and stores the control data of said data in control memory 14 (Step 34). This control data contains the address of the data to be written in the magnetic disk drive 13, the time at which the write command for said data is received, the serial number of the write command, and a pointer for the data in the cache memory 12. The received serial number of the write command is one number out of a series of numbers attached to the write commands received by said disk subsystem 3. Subsequently, the subsystem 3 reports the completion of the process of the write command to the higher device 1 (Step 35). The data stored in cache

memory 12 is later recorded on the magnetic disk drive 13, but this belongs to the prior art, therefore, is not explained in detail in the present invention. (0027)

The disk subsystem controller 17, if it finds that the remote copying is not in a stop status (No response in Step 36) after referring to the control bit in the control memory 14, issues a write command to the disk subsystem 7 connected to it, via the interface controller 11 and interface cable 4, asynchronously with processing of the write command from the higher device 1 (Step 37). In this context, this write command includes, like the write command received from the higher device 1, the command to write said data and the command to transfer said data, but this command further includes the time when said data is received and the serial number attached to the received data. If the remote copying is in a stop status (Yes response in Step 36), the disk subsystem 3 stores the address data at which to write the data in the magnetic disk drive 13 in the control memory 14 of it own disk subsystem 3 (Step 38), and issues the write command to its partner subsystem 7 after the remote copying status returns to a normal status. (0028)

The disk subsystem 7 receives the write command issued from the disk subsystem 3 via the interface controller 11 (Step 41). This write

command includes not only the data but the time when the disk subsystem of the primary center has received the data and the serial number with which the disk subsystem of the primary center has received said data. The disk subsystem 7 begins the process based on the received write command. The disk subsystem controller 17 of the disk subsystem 7 rearranges the received write commands according to a sequence of the serial numbers and checks out if there is a missing number, i.e., a missing write command in the sequence. Subsequently, the disk subsystem 7 stores the received data in cache memory 12 as "tentative data," stores the control data of said data in the control memory 14, and reports the completion of the data reception to its partner disk subsystem 3 (Step 42).

(0029)

Subsequently, the disk subsystem controller 17 of the disk subsystem 7 selects a candidate of the correct time out of the time data contained in the received write commands and determines the correct time candidate in cooperation with the other disk subsystem 7 (Step 43). Determining the candidate for the correct time is to be explained later. When the correct time is thus determined, the data attached with the reception time preceding said time is stored in cache memory 12 as "correct data" (Step 44). The data stored in the cache memory 12 is recorded later on the magnetic disk.

(0030)

Fig. 4 shows a flow of data in the process in which the data attached with the reception time and reception serial number are remote copied. Five sets of data to be written in the higher device 1, D1, D2, D3, D4, and D5, are generated in this order in the higher device 1, and the higher device 1 successively issues the write commands to the disk subsystem 3 – 1 and disk subsystem 3 – 2. The disk subsystem 3 – 1 receives data D1, D3, and D5 and attaches the reception serial number and the reception time to them as S1/T1, S2/T2, and S3/T5 in this order. On the other hand, the disk subsystem 3 – 2 receives data D2 and D4, and attaches the reception serial number and the reception time to them as S1/T2 and S2/T4 in this order. The disk subsystem 3 – 1, after storing the received data in the cache memory 12 in a sequence of the serial numbers, issues the write command of each data to the disk subsystem 7 – 1. On the other hand, the disk subsystem 3 – 2, after storing the received data in the cache memory 12 in a sequence of the serial numbers, issues the write command of each data to the disk subsystem 7 – 2.

(0031)

The disk subsystem 7- 1, upon receiving the data attached with the reception serial number and reception time, stores the data in cache memory

12 as a "tentative data." Subsequently, the data subsystem 7 – 1 and the data subsystem 7 – 2 compare the reception time of the newest data D5 out of the data received by disk subsystem 7-1 and that of the newest data D4 received disk subsystem 7 – 2. These data D5 and D4 are referred to as the correct time candidates. In this example, since the time T4 attached to the data D4 is judged as being older than the time T5 attached to data D5, the time T4 attached to data D4 is determined as the correct time, and the data D2 and data D4 attached with the time preceding this time are copied to cache memory 12 as the "correct data."

(0032)

As a result of above process, data D5 remains as the "tentative data," and data D5 is copied to cache memory 12 later than data D4 is. By preserving the sequential order of updating data D4 and data D5, these data can be recovered and their remote copying can be restarted when the remote copying is interrupted by a disaster. By contrast, if a disaster has occurred after data D5 is updated and copied prior to updating and copying of data D4, the data may be possibly lost in most cases, making it difficult to recover and copy the data. On the other hand, the sequence of updating the data D1 and D3 of disk subsystem 7 – 1 and data D2 and data D4 of disk subsystem 7 – 2, all of which are attached with the correct time, follows the

same sequence of updating data D1 and D3 of disk subsystem 3 – 1 and data D2 and D4 of disk subsystem 3 – 2; thereby preserving the sequence of updating.

(0033)

As explained above, the serial reception numbers indicate the updating sequence in one unit of the disk subsystems 3 and the updating sequence in one unit of the disk subsystems 7. This serial reception numbers are used to check out the missing data. On the other hand, the reception times indicate the time values in the data-updating sequence by multiple units in disk subsystem 3 and are used for the multiple units of the disk subsystem to determine the correct time.

(0034)

Fig. 5 shows steps of the process of disk subsystem 7 in determining the correct time. The disk subsystem controller 17 of disk subsystem 7 that issues a request for determining the correct time out of the correct time candidates selects the candidate in its own device (Step 51); more specifically, it selects the write command attached with the newest reception time out of the write commands arranged in a sequence of the serial reception numbers of data; subsequently, it commands the adjacent disk subsystem 7 connected with interface cable 5 to determine the correct time

candidate (Step 52). This command includes the request indicating the request for determining said candidate, the time attached to the correct time candidate, and the product number of said disk subsystem 7 (product No.). At this time, the time when the request is made is stored in its own control memory 14.

(0035)

The adjacent disk subsystem 7, upon receiving this command, determines whether the product number in the command agrees with the product number of it own device (Step 53); if the product number does not agree with that in its own device (No response in Step 53), it selects the correct time candidate by the above process in its own device (Step 54); subsequently, it compares the time of the received correct time candidate with the time of the correct time candidate in its own device (Step 55). If the time in its own device is older (Yes response in Step 55), the adjacent disk subsystem 7 replaces the determined time of the correct time candidate and the product No. with the time of the correct time candidate and the product No. in its own device (Step 56). If the time in its own device is newer (No response in Step 55), it moves to the step 57; subsequently, it commands the adjacent disk subsystem 7 connected by interface cable 5 to determine the correct time candidate between the replacement time in its own device and

the received correct time candidate (Step 57). It also stores the time of this request in its own control memory 14. After the process in Step 57, it returns to the Step 53.

(0036)

The adjacent disk subsystem 7, upon receiving said command requesting to determine the correct time, determines the correct time candidate, for which the request was made, as the correct time (Step 58) when its product No. in the command agrees with the product No. of its own device (Yes response in Step 58). More specifically, the correct time candidate is selected by every device of the disk subsystem 7, and the oldest time among their candidates is determined as the correct time.

(0037)

Fig. 6 shows a flow of the process by the disk subsystem in notifying the correct time. The disk subsystem 7 that has determined the correct time stores the data attached with the time preceding the correct time as the "correct data" in cache memory 12 (Step 61); subsequently, it issues a command notifying the adjacent disk subsystem 7 of the determined correct time (Step 62). This command contains the request indicating the notice of the correct time, the determined correct time, and the product No. of said

disk subsystem 7. At this time, it also stores the time when the command was issued in the control memory 14.

(0038)

The adjacent disk subsystem 7, receiving this command, determines whether the product No. contained in the command agrees with that of its own device (Step 63); if the product No. does not agree with that of its own device (No response in Step 63), the adjacent disk subsystem 7 stores the data attached with the time preceding to the correct time that was received from disk subsystem 3 in cache memory 12 as the "correct data" (Step 64); subsequently, it sends the received command to the adjacent disk subsystem 7 to notify the correct time (Step 65), and returns to the Step 63.

(0039)

The adjacent disk subsystem 7, upon receiving the command notifying the correct time, deletes (Step 66) from the control memory 14 the unnecessary data, such as the product No. out of the data attached with the time preceding the correct time when the product No. in the command agrees with the product No. of its own device. As a result of the above process, all the write commands to which all the subsystems 7 have attached the time preceding the correct time are provided to the data to be updated in

the disk subsystem 7, and a new command after the correct time remains as the subsequent correct time candidate to be selected.

(0040)

In one method of activating the request for determining the correct time, the disk subsystem to be activated first is preselected, and hereinafter, the disk subsystem 7 that has determined the correct time activates the request for determining the next correct time. Or one unit of the disk subsystems 7 may activate said request periodically. Either method can be used as long as the step of determining the correct time and the determined correct time are notified periodically in the system.

(0041)

In the above example, the manufactured product No. is used as an identifier of each disk subsystem 7, but any identifier can be used as long as it can identify the every disk subsystem 7 of the remote center.

(0042)

The disk subsystem 7, if the correct time determining and requesting command which it has issued did not return or the correct time notice was not received within the prescribed time period, determines that the correct time was not determined or the correct time notice was not completed for some failure, sets the control bit in the control memory 14 to the remote

copy stopping status, and stops the remote copying. When the remote copying cannot be continued because of a failure on the side of disk subsystem 7, it stops the remote copying, and notifies the partner subsystem 3 of said stop of remote copying. The disk subsystem 3, upon receiving this notice, sets the control bit in the control memory 14 to the remote copy stopping status, stops the remote copying, keeps the write command received from the higher device from being transferred to the disk subsystem 7. Also, when the remote copying cannot be continued because of a failure on the side of disk subsystem 3, it likewise stops the remote copying and notifies its partner disk subsystem 7 of said stop of remote copying. Or in the case when the disk subsystem 3 and the disk subsystem 7 have detected the communication unable due to failure of interface cable 4, they likewise stop the remote copying. In addition, at a proper point in time, they can restart the remote copying by changing the control bit in the control memory 14 by the command from the service processor panel 18 of disk subsystem 3 or disk subsystem 7.

(0043)

In addition, a temporary stop and restart of remote copying can be designated by a volume pair unit, which is a minimal control unit of the recording region in the disk subsystem when the remote copying is preset.

The volume pair refers to a pair of volumes formed by a volume in the disk subsystem 3 to be copied from and by a volume in the disk subsystem 7 to be copied to. At least one volume pair is defined as one volume group, and the status can be changed by a volume group unit. The data updating inside the volume group does not affect other volume groups. In this case, the control bit in the control memory 14 is installed per a volume pair or volume group. The disk subsystem 3 refers to the control bit in the control memory 14 that corresponds to the volume or volume group in Step 36 to determine whether the remote copying of the data is stopped or not.

(0044)

Accordingly, in the system wherein all the business data stored in the multiple disk subsystems are constantly duplicated by remote copying, when some data at the operation completion point needs to be used, said data at the operation completion point will be put on hold in the disk subsystem if the remote copying by the volume pair or volume group, in which said operation data has been stored at the operation completion point, is temporarily stopped. Therefore, the operation can be continued at the remote center.

(0045)

Once the temporary stop is released, the disk subsystems 3 and 7 restart the remote copying process and transfer to the disk subsystem 7 the

write command of the data written only in the disk subsystem 3 of the primary center during the temporary stop as well as the data which is not yet defined as the "correct data" in disk subsystem 7. Thus the data is defined as the "correct data" defining process can be restarted. By this, a normal remote copying status is resumed.

(0046)

In the system structure shown in the above embodiment example, the disk subsystem 3 of the primary center 9 and the disk subsystem 7 of the remote center 10 are connected in one-on-one relationship. As the other embodiment example, it is also possible to construct the system wherein multiple disk subsystems 3 are connected to one unit out of disk subsystems 7, and remote copying is performed between this disk subsystem 7 and said multiple disk subsystems 3. In this structure, the disk subsystem 7, upon receiving data in Step 41, rearranges the received write commands in a sequence of the serial numbers of the received data by using the product numbers of the disk subsystems, selects the newest time out of the times attached to the write commands of the disk subsystems 3, compares the times selected by the disk subsystems 3, selects the oldest time out of them as the correct time candidate, and determines the correct time.

(0047)

Moreover, as another embodiment example, it is possible to construct the system structure, wherein one unit out of the disk subsystems 7 of the remote center is used as a master disk subsystem (generally referred to as a master memory subsystem), and all the disk subsystems 7 other than the master disk subsystem in the remote center 10 are connected to the master subsystem. In this structure, the master disk subsystem acquires by inquiry the correct time candidate retained in the other disk subsystems. Subsequently, the master disk subsystem compares the acquired correct time candidate with the its own correct time candidate, determines the oldest time as the correct time, and notifies the other disk subsystems of the remote center of the determined correct time; thus, each disk subsystem 7 determines the "correct data."

(0048)

In this embodiment example, since the disk subsystem 7 of the remote center 10 determines the correct data based on the correct time candidate and the correct time, asynchronous remote copying can be embodied by the disk subsystems while preserving the sequence of data updating without inserting the higher device, without periodically interrupting the remote copying, without installing a gateway subsystem for preserving the data updating

sequence on the primary center 9 and remote center 10, and without reducing the process performance of disk subsystem 3 of the primary center 9.

(0049)

According to the embodiment examples of the present invention, the data updating sequence is ensured by using the time acquired from the clock contained in the disk subsystem, so asynchronous remote copying can be embodied while preserving the sequence of data updating without inserting the higher device 1, even in the system wherein the time data is not attached to the data transferred from the higher device 1 to the disk subsystem 3.

(0050)

In addition, even if the function of the primary center is broken down because of a disaster or failure of the primary center, data for which the data updating sequence is preserved is stored in the disk subsystem 7 of the remote center 10. These processes are carried out only by the functions of the disk subsystems 3 and 7 without imposing a burden on the processing capability of the higher device 1. In the case when a disaster falls on the primary center, a recovery operation such as restarting of a job, can be carried out by using the data in the disk subsystem 7, so the operation can be restarted.

(0051)

(Advantage of the Invention)

As explained above, an asynchronous remote copying system, which can be easily introduced and does not reduce the processing performance of a host computer of a primary center, can be implemented only by the functions of memory subsystems, while ensuring the data updating sequence to the extent that the users expect, without introducing a new software application into the higher device.

## BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 shows the entire structure of the remote copying system as one embodiment example of the present invention.

Fig. 2 shows the internal structure of the disk subsystem in the embodiment example.

Fig. 3 shows a flow of the steps of data duplication between the disk subsystem 3 and the disk subsystem 7 in the embodiment example.

Fig. 4 illustrates one example of a data flow in which the data, each being attached with a reception time and a serial reception number, are remote copied.

Fig. 5 shows the steps of a process in which the disk subsystem 7 determines the correct time in the embodiment example.

Fig. 6 shows the steps of a process in which the disk subsystem 7 notifies the correct time in the embodiment example.

1. higher device
3. disk subsystem
4. interface cable
5. interface cable
7. disk subsystem
9. primary center
10. remote center
15. clock
16. time correction section